Deciphering the language of fungal pathogen recognition receptors

Witold Dyrka, Pascal Durrens, Sven J Saupe, Mathieu Paoletti, David James Sherman

Deciphering the language of fungal pathogen recognition receptors

Witold Dyrka[1,2,3*], Pascal Durrens[1], Sven J. Saupe[2], Mathieu Paoletti[2#] and David J. Sherman[1$]

[1] INRIA-Université Bordeaux-CNRS, Team MAGNOME, Talence, France,
[2] Institut de Biochimie et de Génétique Cellulaire, CNRS-Université de Bordeaux, France,
[3] Department of Biomedical Engineering, Wroclaw University of Technology, Poland

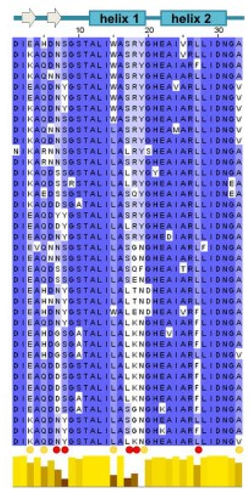
*witold.dyrka@pwr.edu.pl #mathieu.paoletti@ibgc.cnrs.fr $david.sherman@inria.fr

Overview

The NLR family of receptors plays a key role in the innate immune system of animals, plants and fungi. In the latter two phyla NLRs adapt quickly to ever-changing pathogen-specific invasion markers thanks to their repeat-based architecture, which can produce diversity of recognition epitopes through unequal crossing-over and mutation. Characterizing computationally the language of these pathogen recognition receptors can provide insight into the molecular mechanisms of immune response and describe the limits of the pathogen targets that can be recognized. In this work, we model generation and selection of the recognition paratopes as a stochastic string rewriting system with constraints, tuned by analysis of observed evolutionary processes, and validated with regard to a large dataset of fungal NLRs. The methodology developed in this work is general and therefore can be applied to any class of amino acid repeats generated by unequal crossing-over for which an equivalent high quality dataset is available.

1. Fungi are genuine interactors which have to deal constantly with multiple hostile non-self. It has been proposed that their ultimate line of defense is a programmed cell death triggered by recognition of pathogen effectors or their markers.

2. We hypothesized that fungi recognize the invasion markers using the repeat domain of NLR proteins. The repeats are often highly conserved internally in each sequence, which allows for their fast rearrangement through the unequal crossing-over, a process up to 100,000 times quicker than the standard mutation.



Computational model of repeats rearrangement

Stochastic string rewriting system with constraints $\langle \Sigma, R, P, Q \rangle$

Σ – alphabet of 20 amino acid types

R – set of rewriting rules $u \rightarrow v \in \Sigma^*$

P – set of rule probabilities

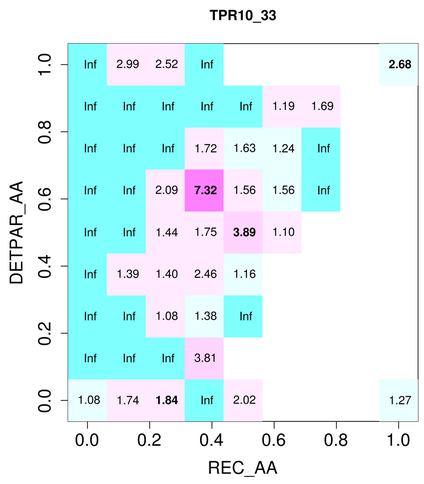
Q – set of constraints, e.g.

allowed positions and lengths of crossing-overs

external constraints acting on repeats („selective pressure")

3. In each family of NLR repeats (Ankyrin, TPR and WD40), we identified several positions which are highly variable despite overall high conservation of repeats. These positions, often found to be under positive selection, are expected to form the recognition paratopes quickly adapting to fast-evolving pathogens.

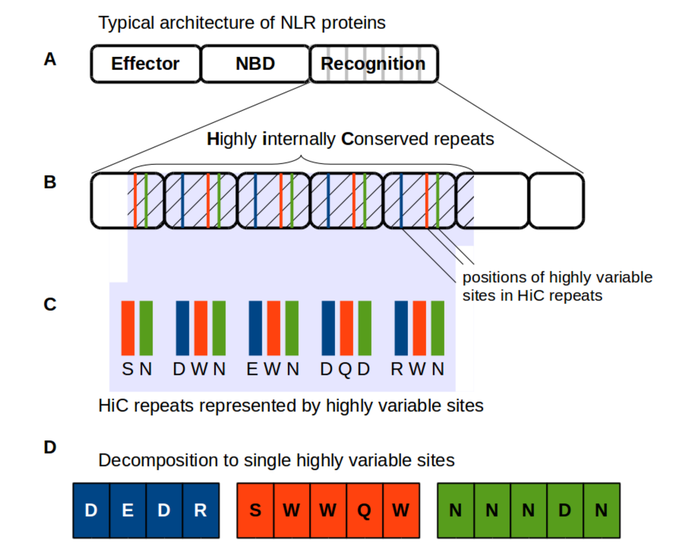
6. The approach also allows exploring solution space in order to find discrepancies between real and simulated data. In a preliminary study, we found a significantly overrepresented pattern at one position in the TPR family:
R-[SYQFW](1,3)-R.



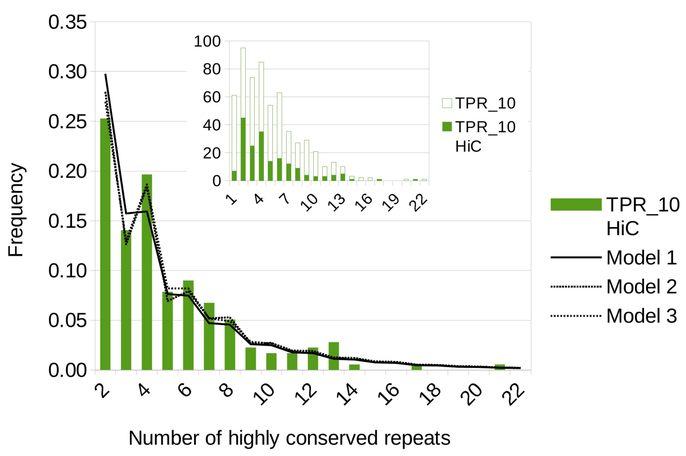
Key properties of the model

It is easy to show that
for realistic parametrization of the crossing-over and mutation and simple constraints
the system generates a single stationary distribution of:
amino-acid composition,
repeat number,
repeat sequence.

Therefore, differences between distributions
generated by the model
and observed in the reality
can be interpreted as an effect of external pressures.



5. The model explained the even-odd periodicity observed in the repeat number distribution of the TPR family of receptors. Moreover, in comparison to the simulated data, the amino-acid composition of real sequences revealed preferences consistent with the putative role of interacting paratope (bias towards polar residues, tyrosine and often tryptophan)



4. Repeat regions in NLR were decomposed to sequences of aminoacids at single highly variable sites. We simulated evolution of amino-acid sequences at each site using our stochastic string rewriting system with constraints, and compared results to real data consisting of 550 sequences.

References: Paoletti & Saupe (2009). Bioessays 31:1201.
Chevanne *et al.* (2010) BMC Evolutionary Biology 10:134.
Dyrka *et al.* (2014), Genome Biology and Evolution 6:3137.

This research was partially funded by ANR-11-BSV3-0019